Supramolecular Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713649759

An improved method for determining bimolecular association constants from NMR titration experiments

Richard E. Barrans Jr.[a]; Dennis A. Dougherty[a]
[a] Contribution No. 8886 from the Arnold and Mabel Beckman Laboratories of Chemical Synthesis California Institute of Technology, Pasadena, California

To cite this Article Barrans Jr., Richard E. and Dougherty, Dennis A.(1994) 'An improved method for determining bimolecular association constants from NMR titration experiments', Supramolecular Chemistry, 4: 2, 121 — 130
To link to this Article: DOI: 10.1080/10610279408029871
URL: http://dx.doi.org/10.1080/10610279408029871

SUPRAMOLECULAR CHEMISTRY, Vol. 4. pp. 121–130
Reprints available directly from the publisher
Photocopying permitted by license only

An improved method for determining bimolecular association constants from NMR titration experiments

RICHARD E. BARRANS, JR. and DENNIS A. DOUGHERTY*

Contribution No. 8886 from the Arnold and Mabel Beckman Laboratories of Chemical Synthesis California Institute of Technology, Pasadena, California 91125

(Received: November 6, 1993)

A nonlinear regression procedure for fitting estimates of an association constant and saturation shifts to NMR titration experiments under fast-exchange conditions is described. The method assigns weights to each observation by propagating measurement errors through the fitted model. A series of Monte Carlo studies simulating a variety of possible experimental conditions has shown this method to be significantly superior to other methods commonly in use.

I. INTRODUCTION

A primary goal of research in molecular recognition is development of an understanding of how and why intermolecular association processes occur. Studies of small (compared to biological structures) model systems allow *quantitative* comparisons of related systems, and thus the possibility of uncovering the fundamental natures of the interactions responsible for noncovalent binding.

The most common approach to obtaining quantitative binding data has been the NMR titration approach. If complexation induces significant chemical shift changes in the guest and/or host, NMR provides an ideal probe of binding. In the overwhelming majority of cases, complexation/decomplexation is fast on the NMR timescale. The observed chemical shift then is a weighted average of free and bound shifts, where the weighting factor F is the fraction of the species bound (eq. 1). The fraction F depends on the association constant K and the total concentrations of host and guest, $[H]_0$ and $[G]_0$. Equation 1 readily rearranges to equation 2, where $D \equiv \delta_{free} - \delta_{bound}$. The equilibrium relation

$$K = \frac{[H{\bullet}G]}{[H]_0[G]_0}$$

and the mass balance relations $[H]_0 = [H{\bullet}G] + [H]$ and $[G]_0 = [H]{\bullet}G] + [G]$ combine to give the expression for F (eq. 3).

$$\delta_{obs} = \delta_{free}(1 - F) + \delta_{bound}F \quad (1)$$

$$\delta_{obs} = \delta_{free} - DF \quad (2)$$

$$F = \frac{1}{2[G]_0}\left\{[H]_0 + [G]_0 + \frac{1}{K} - \sqrt{\left([H]_0 + [G]_0 + \frac{1}{K}\right)^2 - 4[H]_0[G]_0}\right\} \quad (3)$$

By varying the concentrations and thus the fraction bound, one can produce a "titration" series. A nonlinear fit of the data then determines the two unknown parameters: the association constant, K; and the saturation chemical shift of the guest D.[1] Many discussions of the different ways to perform such an experiment and to extract the K and D values have been given, and general guidelines are available.[2,3]

We have been interested for some time in two aspects of this problem.

1) What is the *optimal* way to obtain K and D from an NMR experiment?

2) What are meaningful error bars for these derived quantities?

In the present work we will address the first question. We will make use of a suite of programs designed to evaluate the statistical significance of many aspects of

*Corresponding author:

the NMR titration problem — and thus question 2 — but the details of these programs will be provided elsewhere.[4,5]

For the present purposes we shall assume the reader has some familiarity with the NMR technique as applied to molecular recognition problems. We also assume that 1:1 binding has been established, and only this type of binding is considered. Of course, in actual experiments there are often complications due to 1:2 or 2:1 binding. Also, it is worth emphasizing here, as has often been done before,[1,2] that, in an actual experiment, the investigator should be sure that the species being observed undergo a large change in the % bound during the course of the experiment. These issues of experimental design are, of course, important, but are not the issue here. Here we investigate which is the best way to evaluate a data set once it is acquired.

The quantities D and K cannot be directly measured. However, estimates of D and K may be evaluated by substituting them, along with measured values of the explanatory variables $[H]_0$, $[G]_0$ and δ_{free}, into eqs. 2 and 3. This leads to δ_{calc}, the predicted value of the response variable δ_{obs}. If the model is correct, if the measurements are performed without error, and if the estimates of K and D are equal to the true values of these parameters, then the predicted δ_{calc} and measured δ_{obs} will be identical. Measurement errors, however, will make the observations deviate from their predicted values, even if the model and parameters are correct.

Least-squares estimation.

Nonetheless, a good model should duplicate the observed data as closely as possible. This is reflected in the criteria used for finding parameter estimates. In least-squares estimation, the parameters are adjusted to minimize the unweighted sum of squared residuals (SSR). For an experiment involving observations of P protons in each of N samples, this score is given in eq. 4

$$\mathrm{SSR} = \sum_{p=1}^{P} \sum_{i=1}^{N} (\delta_{\mathrm{calc}\ pi} - \delta_{\mathrm{obs}\ pi})^2 \qquad (4)$$

Minimization of SSR is not the only useful criterion for evaluating parameter estimates. The real objective of parameter estimation is not merely to find a model to fit some data set, but to find the true parameter values. To that end, if a number of experiments are performed to evaluate K and D of some host/guest system, a good estimation procedure should determine values from all of these experiments that are narrowly distributed about the true values. An estimator may thus be evaluated by its *bias* and *variance*. Bias is the difference between the expectation, or mean, of the estimator and the true value of the parameter. An ideal estimator is unbiased: its expectation is exactly the true parameter value. The variance of a random variable is the average square of the difference between a given occurrence of the variable and the variable's true mean. This is a measure of the spread of the variable: a small variance means that the variable's distribution is very compact. A good estimator thus has a small bias and a small variance.

The popularity of least squares arises from several considerations. In the first place, it is familiar and convenient. Furthermore, it has optimal properties in certain cases.[6] In particular, if the predictive model relating the explanatory variables to the response variables is a linear function of the adjustable parameters, finding the parameters that minimize SSR is a simple matter of solving a system of simultaneous linear equations. The most attention has been given to cases in which measurement errors are in only the response variables, with the explanatory variables being known exactly. Least squares estimators are then unbiased, and if the measurement errors are normally distributed, they have the smallest variance of any possible unbiased estimator. Even if the predictive model is a nonlinear function of the parameters, having measurement error in only the response variables still makes least squares consistent, that is, unbiased in the limit of infinite sample size.

Such conditions are not fulfilled by the model of equation 2, which is a nonlinear function of the parameter K. In addition, measurement errors occur in all of the measured variables, not only in the response variables.[7] As a result, there is no guarantee that least-squares parameter estimates are good by any standards. In addition, statistical theory is unable to *a priori* identify estimators that have optimal properties for this model. Consequently, we have undertaken a qualitative, empirical approach to the problem.

II. FITTING METHODS

A. Standard Approach.

In most NMR titration experiments, the parameters K and D are assigned by minimizing the unweighted SSR of equation 5.

$$\mathrm{SSR} = \sum_{i=1}^{N} (\delta_{\mathrm{calc}\ pi} - \delta_{\mathrm{obs}\ pi})^2 \qquad (5)$$

This is simpler than the SSR of equation 4; it counts only the observations of a single proton. If a binding study is performed in which the resonances of more than one proton are followed, K and D are determined separately for each proton. In principle, the observations of all protons should give the same estimate of K. In fact, measurement errors ensure that this will never occur exactly. When a binding study of a system with P observed protons is carried out, P different estimates of the associ-

ation constant are returned. A "best" estimate of the association constant is typically devised by averaging the estimates from the individual protons, or by disregarding all protons but the one the model was best able to fit.

B. MULTIFIT.

As a first step up from this simple analysis, we developed a Pascal program (MULTIFIT)[8] to estimate a single binding constant for all of the observations in a binding study. If P protons are observed in a study, there are $P+1$ adjustable parameters in the predictive model: K and the P D's. The association constant K returned by this procedure is the true least squares estimate. It is more reliable than an estimate from any single proton, because it is based on more information. Furthermore, the D values returned are consistent with each other, which is not necessarily the case if the protons are fitted separately. This allows confident comparison of the D values of different protons. Such an ability is essential for describing the geometry of a complex.

This procedure, nonetheless, has some drawbacks. The most severe is its assignment of equal weights to all observations. For instance, it is common for different protons to have values of D that are very different in magnitude. Protons with large absolute values of D will change their peak positions in an NMR binding study much more significantly than will protons with smaller D's. As a result, an unweighted least-squares estimation procedure places the greatest relative importance on fitting the observations of the proton whose signals move the most. This is not the best use of all of the information available in the experiment.

C. EMUL

1. Design.

To combat this drawback, we have developed a *weighted* least-squares fitting program (EMUL).[8] This program minimizes SSR* (eq. 6), a different loss score than the SSR of MULTIFIT.

$$\mathrm{SSR}^* = \sum_{p=1}^{P} \sum_{i=1}^{N} \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})^2}{\sigma_{pi}^2} \qquad (6)$$

In this equation, σ_{pi} is the estimated inaccuracy in predicting the resonance of proton p in sample i. This value is determined by a first-order approximation of the influence of each measurement error on the eventual magnitude of the residual $\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}$. If this residual is affected by L measured variables χ_j, each of which can be thought of as a random variable with variance σ_{xj}^2, then the estimate of σ_{pi}^2 is given in eq. 7.[9]

$$\sigma_{pi}^2 = \sum_{j=1}^{L} \sigma_{xj}^2 \left(\frac{\partial(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\partial \chi_j} \right)^2 \qquad (7)$$

2. Execution.

The first step toward creating a procedure to minimize SSR* was to develop the means to evaluate SSR* itself. This required identifying the fundamental random variables χ_j, their uncertainties $\sigma_{\chi j}$, and the derivatives

$$\frac{\partial(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\partial \chi_j}.$$

The identities of the fundamental random variables of an experiment depend on the design of the experiment itself. Strictly, every measurement performed is a random variable. Binding studies are typically performed by combining stock solutions of host and guest together with additional solvent in an NMR sample tube, and recording the spectrum. The values of $[H]_0$ and $[G]_0$ are then altered by adding more host solution, guest solution, or solvent, and the NMR spectrum is again recorded. The steps of adding solution and recording the spectrum are repeated several times, and the spectra of uncomplexed host and guest are measured independently.

The fundamental random variables contributing to a single observation are:

- The host and guest concentrations, $[H]_S$ and $[G]_S$, of every stock solution used to make up the sample,
- the volume V_a of each solution aliquot added to the sample tube,
- The calibration I of the delivery devices (pipets or syringes) employed to add the aliquots, and
- the NMR peak position measurements $\delta_{\mathrm{obs}\,pi}$ and $\delta_{\mathrm{free}\,p}$

Uncertainties in stock solution concentrations will vary considerably from system to system, depending on the amount of material available, the method of concentration measurement, etc. In our experience the uncertainties here can be large — perhaps as much as ± 5%. Aliquot volumes are determined by two related but independent random variables: the precision and accuracy of delivery devices. Precision is the reproducibility of volumes added by a device. Accuracy is a measure of the likely calibration error of the delivery device. The volumes of all aliquots delivered by a single device will be mis-measured by the same proportional amount; for instance, they all may be 2% too low. Thus, the difference between an aliquot's true and measured volumes is (measured value) × (calibration error) + (reproducibility error). The final fundamental random variables considered are the NMR measurements. Because NMR signals

Table I Names of experimental designs in Monte Carlo studies.

	$K =_a$			
design	10^3	10^4	10^5	10^6
(1) adding host	H3	H4	H5	H6
(2) adding guest	G3	G4	G5	G6
(3) adding diluent	D3	D4	D5	D6
(4) continuous variation (Job)	J3*	J4	J5	J6
(5) constant $[H]_0$	V3	V4	V5	V6

[a]In M^{-1}. *Not Included in simulations.

are well-resolved and reproducible, the errors in these variables are generally small. The principal source of such error is the digitization of the spectrum: the peak position cannot be known more specifically than the distance between two points. Another possible contributor to peak position measurement error is peak width. If a peak is very broad, as often happens in exchanging systems such as these, it can be difficult to tell exactly where its center lies.

Once the fundamental random variables have been identified, it is necessary to determine their impacts upon the observations according to equation 7. This task is tedious but straightforward: it requires only differentiation of $(\delta_{\text{calc } pi} - \delta_{\text{obs } pi})$ with respect to each of the fundamental variables. The details of determining these derivatives are given elsewhere.[5]

III. COMPARISON OF FITTING METHODS

We will now compare the performance of a number of methods of data fitting, including the approaches described here. First we will evaluate EMUL, and then a variety of other procedures in current use

A. Design.

In order to compare different fitting schemes to each other, we have tested them on artificial data sets generated by Monte Carlo simulation experiments.[10] The goal is to generate a representative series of "experiments", each of which can be addressed using each of the methods of interest. As each data set is fitted by the regression procedures being compared, parameter estimates from each procedure are created. The behavior of the estimates over a large number of data sets provides an empirical basis for the comparison of the different procedures.

Such comparisons were carried out for a variety of experimental designs, covering the range of binding constant values that can reasonably be determined from NMR titration experiments. Five basic types of experimental design were modeled: (1) adding aliquots of host stock solution to a sample tube containing guest; (2) adding aliquots of guest stock solution to a sample tube containing host; (3) adding aliquots of diluent to a sample tube containing both host and guest; (4) a Job or continuous variation study, in which $([H]_0 + [G]_0)$ is the same in all samples, and the mole fraction of each species is varied in equal steps from 0 to 1; and (5) making $[H]_0$ the same in all samples, changing only the concentration of guest. Each experiment involved fifteen observed samples; in each of these samples $[H]_0$ was between 10 and 200 μM, and $[G]_0$ was between 10 and 500 μM. Two proton signals were followed, one from the host and the other from the guest. D of the host proton was -100 Hz, and D of the guest proton was +500 Hz. For each experimental design, four association constants K were considered: 10^3, 10^4, 10^5, and 10^6 M^{-1}. Each of these twenty experiments was designed to provide a good measure of the association constant by keeping the fraction of the minor component bound between 0.2 and 0.8.[3] These sets are summarized in Table I. When the association constant was 10^3, the method of continuous variation (design 4) proved to be an extremely poor experimental design. Small simulated measurement errors led to a preponderance of terrible parameter estimates. As a result, this set was not included in the large study; only the remaining nineteen sets were used.

Except as specified otherwise, measurement errors followed normal distributions as follows. The standard deviation of stock solution concentration measurements was 5%, and the standard deviation of NMR peak position measurements was 0.5 Hz. Aliquot volume errors and delivery device calibration errors depended on the delivery device used. Delivery device accuracy and precision error distributions were adapted from the specifications for Eppendorf Varipette 4810 piston stroke pipettes.

B. Testing Error Propagation.

The first comparisons performed were intended to assess the importance of propagating the measurement errors in each of the fundamental explanatory variables. In each of these comparisons, 1500 Monte Carlo repetitions of each of the nineteen experiments under consideration were performed. The data set from each repetition was subjected to three types of least-squares fit. The first method minimized the sum of squares of the unweighted residuals, SSR (equation 4). The third minimized the

sum of squares of the weighted residuals, SSR* (equation 6), with the weights calculated by propagation of all measurement errors according to equation 7. The second also minimized a weighted sum of squares, but with the errors in one type of measurement not propagated. This was to determine if propagating each of the different types of measurement error was beneficial or detrimental. If it is disadvantageous to propagate a certain type of measurement error, then this abbreviated procedure should perform better than the one with full propagation.

C. Evaluating Performance.

Six measures of performance of the fitting methods were calculated under each experimental condition. These measures were the medians and standard deviations of the three fitted parameters K, D_1, and D_2. These provide a way to evaluate the bias and variance of the parameter estimates from the fitting procedures. Medians were evaluated in preference to means because the median is a more robust measure of central tendency. The performances of the three fitting procedures with respect to each of these six measures were compared and ranked. The method with the best performance in a measure received the rank of 1, the second best received the rank of 2, and the worst the rank of 3. If some procedures performed indistinguishably well (if they were tied), each received the same rank, which was the average of the ranks they would have received if they had been slightly different.

Consider as an example the replications of experiment H3, in which the association constant is 10^3 M^{-1} and the protocol follows experimental design 1. The distribution of the estimates of K from the unweighted minimization had a standard deviation of 138; from both the fully- and partially-weighted procedures, the standard deviation of this same estimate was 149. Thus, the unweighted procedure received a rank of 1, and the others both received ranks of 2.5. The *medians* of all three of these distributions were 1.00×10^3 M^{-1}, however, so each procedure received a rank of 2 for this measure.

Each study thus produced $19 \times 6 = 114$ sets of rankings. In order to determine if one fitting procedure performed significantly better overall than any of the others, these rankings were evaluated by a Friedman-Cochran-McNemar test.[11] This nonparametric statistical test is designed to determine if there is a significant difference between s subjects that have been ranked by N independent judges. Let us define the total rank R_i of the ith subject as the sum of the N ranks received by that subject.

$$R_i = \sum_{j=1}^{N} \text{rank}_{ij}$$

In our case, some of the rank sums R_i contain tie scores. The test statistic is then

$$Q^* = \frac{\frac{12}{Ns(s+1)} \sum_{i=1}^{s} R_i^2 - 3N(s+1)}{1 - \frac{\sum_{j=1}^{N} \sum_{k=1}^{e_j} (d_{kj}^3 - d_{kj})}{Ns(s+1)}}$$

The index d_{kj} here is the number of subjects assigned rank k by judge j. Each judge assigns e_j distinct ranks. If there are no ties, then $e_j = s$; if some subjects are tied, then $1 \leq e_j < s$. If there is no difference between subjects, ranks will be assigned randomly and uniformly, so that all rank sums R_i will be similar. The statistic Q^* will then be distributed as a χ^2 variable with s-1 degrees of freedom. A very large Q^* rejects the null hypothesis that the subjects are indistinguishable.

In this evaluation of fitting methods, the subjects are the fitting methods and the judges are the sets of experimental conditions. It is most informative to make paired comparisons of fitting methods, that is, to compare one method to one other. With three fitting methods, there are

$$\binom{3}{2} = 3$$

such comparisons to make. These comparisons can be carried out by the Friedman-Cochran-McNemar test, with $s = 2$. Since s is 2 instead of 3, slightly different ranks from those assigned from the full set of three fitting methods must be used. These new ranks are easily derived from the old ranks: the subject with the lowest rank is assigned a new rank of 1, and the subject with the highest rank is assigned a new rank of 2. If the subjects are tied, both receive new ranks of 1.5. These ranks determine Q^*. If the two subjects are indistinguishable, Q^* will follow the

$$\chi^2_1$$

distribution. The null hypothesis of indistinguishability is rejected if Q^* falls above some cutoff for this distribution. The 95% cutoff for this distribution, for instance, is 3.84.

The performance of these three fitting methods according to the six different measures can be conveniently summarized in the following manner. If one method performs significantly better than another, that is, if Q^* from a head-to-head comparison is greater than 3.84, then the winning method receives a score of +1 and the losing method receives a score of -1. If no significant difference is found between the two methods, each receives a score of 0. The scores received by a method in its comparisons to the other two methods are added together to give a total score for that measure.

For example, let us examine the standard deviations of K estimates in the test of propagating aliquot volume reproducibility errors. In this comparison, full propagation

Table II Relative performances of fitting methods in a test of error propagation.

	K		D_1		D_2		
error propagation	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
			aliquot volume errors				
none	0	-2	0	-2	0	-2	-6
partial	0	0	0	1	0	1	2
full	0	2	0	1	0	1	4
			device calibration errors				
none	0	-2	-2	-2	0	-2	-8
partial	0	1	1	1	0	1	4
full	0	1	1	1	0	1	4
			stock solution concentration errors				
none	0	-1	-1	-1	0	-1	-4
partial	0	0	0	0	0	-1	-1
full	0	1	1	1	0	2	5
			all NMR spectrometer errors				
none	0	-1	0	0	0	-1	-2
partial	-1	0	0	0	-1	0	-2
full	1	1	0	0	1	1	4
			δ_{free} errors only				
none	0	-2	0	0	0	-2	-4
partial	0	1	0	0	0	1	2
full	0	1	0	0	0	1	2
			δ_{obs} errors only				
none	1	-2	1	0	0	0	0
partial	-2	1	-2	0	0	0	-3
full	1	1	1	0	0	0	3

of errors proved to be significantly superior both to no error propagation at all and to propagation of all errors except for aliquot volume reproducibility errors. Furthermore, the partial propagation method was significantly better than the method of no propagation at all. The scores assigned are thus 1+1 = 2 to the full propagation method, 1-1 = 0 to the partial propagation method, and -1-1 = -2 to the no propagation method.

Table II shows the total scores given to the three fitting methods for each of the six measures of fitting method performance. The final column gives the sum of scores assigned by these six measures for each of the fitting methods. Comparison of the total scores for the competing methods reveals which method performs best overall.

Six separate studies are summarized in Table II. In all of these studies, the fitting method that does not propagate errors performs worse overall than the method employing full error propagation. In no case does propagation of a subset of the measurement errors perform better overall than does full propagation. Consequently, we believe that the full error-propagation method is justified. There is no indication that propagating fewer measurement errors would produce a better estimation procedure.

D. Other Fitting Procedures.

A similar series of Monte Carlo studies was also performed to compare a larger class of fitting procedures. In these studies, five fitting methods were compared. The first of these methods is a currently popular method for finding K from NMR titration experiments, developed independently by Creswell and Allred[12] and by Horman and Dreux.[13] This method involves using δ_{free} as an adjustable parameter instead of an independently-measured variable. Conceivably, this approach could have an advantage over methods in which δ_{free} is directly measured, because the parameter estimates are unaffected by errors in the determination of δ_{free}. In other methods (such as EMUL), if δ_{free} is measured erroneously, the model is systematically compromised. The method of Creswell and Allred determines δ_{free} from the entire data set, instead of relying on a single measurement. Thus, method (I) involves fitting the entire data set at once by adjusting a single association constant, and the free chemical shift and saturation shift ($\delta_{free\ p}$ and D_p) for each proton. All observations are weighted equally, and no use is made of an independent measurement of a proton's free chemical shift. When P protons are observed, this method has $2P+1$ adjustable parameters. As commonly implemented, this method is applied to each proton separately, producing K_p, $\delta_{free\ p}$, and D_p (3P adjustable parameters). However, we have used the MULTIFIT approach here.

We also considered four other methods. (II) Fitting the entire data set by adjusting the same parameters as in method I, but treating an independent measurement of a proton's free chemical shift δ_{free} as an additional observation. This adds one squared residual term ($\delta_{free\ calc\ p}$ –

$\delta_{\text{free obs } p})^2$ to fit the score SSR for every proton observed. This method is intermediate between methods I and IV. (III) Fitting the observations for each proton separately. The association constant K and saturation shift D_p are optimized for the observations on a single proton, and this process is carried out for each proton. After all observations have been modeled in this manner, the estimates of K from each proton are averaged to give the overall "best" estimate of K. This is the standard method described in section IIA. (IV) Fitting the entire data set by adjusting a single association constant K and the saturation shifts D_p of all observed protons. All observations are weighted equally. This procedure has $P+1$ adjustable parameters, and is the method employed by MULTIFIT. (V) Fitting the entire data set by adjusting the association constant K and the saturation shifts D_p of each observed proton. Observations are assigned weights by propagating all measurement errors according to equation 7. This is the method used by EMUL.

These studies were carried out in a manner similar to that used for testing the propagation of the different classes of measurement errors. Each of the 1500 Monte Carlo replications of each of the nineteen experiments was fitted by each of the five fitting methods. The performances of the fitting methods according to the fitted parameter medians and standard deviations were evaluated and ranked for each set of experimental conditions. Head-to-head comparisons of pairs of fitting methods were evaluated by using the Friedman-Cochran-McNemar Q^* statistic, based on the relative ranks of the two compared methods. Since five subjects were evaluated, there were

$$\binom{5}{2}=10$$

pairwise comparisons for each performance measure. Each method received a score of 1, 0, or -1 from each of its pairwise comparisons, which were added together to give a total score for the method. These total scores, and the sums of these scores over the six performance measures, are reported in Table III.

Clearly, the superior method for fitting NMR data under the experimental conditions considered is V, the method that assigns weights by propagating measurement errors. On the other hand, I, which eschews experimental measurement of uncomplexed chemical shifts, is the poorest performer.

The performance of method I would probably improve if the experiments were designed to sample the entire range of chemical shift values for all the protons observed. This would require the fraction bound of each species to range from near zero to near unity in each study. Experimental conditions often prohibit such observations if one is unwilling to measure the spectra of host and guest individually. For example, if the association constant of a given host/guest pair is 10^6 M^{-1}, both species are 80% bound if the total concentration of each is 20 μM. It is often not practical to reduce the fractions bound by making the sample more dilute, because NMR is not sensitive enough to detect lower concentrations. Raising the concentration of one species so that it swamps the other would allow observation of the major component in the almost entirely free state, and of the minor component in the almost entirely bound state. Such an observation is informative for determining δ_{free} of the major component and δ_{bound} of the minor component, but it contains practically no information about the association constant.[3]

We have also evaluated the extent to which the various fitting methods are "robust", this is, able to treat data sets that contain unusually large experimental errors. For example, we have modeled a case in which two independent stock solutions of the same intended concentration were used for one of the species. In the experimental protocol in which host stock solution aliquots are added to the sample, for instance, every other such aliquot was taken from the second stock solution. This design was contrived to test the performance of the fitting methods when the stock solution concentration behaves more like a random error and less like a systematic error. All the experimental designs were perturbed in this fashion, except for the design in which aliquots of diluent are added to the sample. The number of judges in this study, N, was therefore 15 instead of 19. The outcomes of the head-to-head comparisons between fitting methods are presented in Table IV.

Method V is again the best performer under these experimental conditions. Method I is no longer the worst performer; it has been eclipsed by method III, in which separate estimates of K from the individual protons are

Table II. Relative performances of fitting methods.

	K		D_1		D_2		
method	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
I	0	-2	-1	-2	0	-2	-7
II	0	1	0	1	0	0	2
III	0	0	0	-3	0	0	-3
IV	0	-1	0	0	0	-1	-2
V	0	2	1	4	0	3	10

Table IV Relative performances of fitting methods when a duplicate stock solution is used.

	K		D_1		D_2		
method	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
I	0	-1	0	-2	0	-1	-4
II	0	-1	0	1	0	-1	-1
III	0	-1	0	-3	0	-1	-5
IV	0	-1	0	0	0	-1	-2
V	0	4	0	4	0	4	12

Table V Relative performances of fitting methods when the standard deviation of the second observation for each proton is 20 Hz.

	K		D_1		D_2		
method	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
I	-1	-2	-4	-3	-4	-3	-17
II	1	1	1	1	1	0	5
III	-3	-2	1	-3	1	0	-6
IV	1	-1	1	2	1	-1	3
V	2	4	1	3	1	4	15

Table VI Relative performances of fitting methods when the standard deviation of every δ_{free} is 20 Hz.

	K		D_1		D_2		
method	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
I	1	-1	1	-3	1	-1	-2
II	1	0	1	2	2	0	6
III	1	-1	1	-2	1	1	1
IV	1	-1	1	2	0	0	3
V	-4	3	-4	1	-4	0	-8

Table VII Relative performances of fitting methods when the standard deviation of δ_{free} of only the host proton is 20 Hz.

	K		D_1		D_2		
method	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
I	0	0	0	-2	0	-1	-3
II	2	0	0	2	0	0	4
III	-1	0	0	-2	0	1	-2
IV	0	0	0	1	0	0	1
V	-1	0	0	1	0	0	0

averaged to give the overall estimate. Apparently, this method is more vulnerable than the others to vagaries in the stock solution concentrations. This difference may also be a random fluctuation: visual inspection of the five methods under this set of experimental conditions and the set summarized in Table III does not reveal any qualitative differences between these two sets.

The effect of imprecise NMR measurements was investigated in a series of studies. Table V summarizes the results from a study considering a single bad spectrum. In every experiment in this study, the observations of both the host and guest protons have a standard deviation of 20 Hz in the observations of the second sample. As may be expected, method V performs exceedingly well under these conditions.

Another study investigated the effect of extremely imprecise measurements of δ_{free} of both protons. In this study, the standard deviation of these measurements was 20 Hz. Since method I does not use these measurements, one might expect it to perform well under such conditions. The outcome of this study is summarized in Table VI. In this instance, the performance of method V is the worst. Although the large measurement errors in δ_{free} were propagated to assign weights to the observations, this method was unable to obtain good parameter estimates. Like methods III and IV, it has only one opportunity to estimate δ_{free}, and that is in the measurement itself. If the measurement is bad, so is the estimate of δ_{free} and no subsequent observations can improve it. Still, the inferior performance of this method in comparison to methods II and III indicates that the propagation of errors is in fact detrimental to the parameter estimation when δ_{free} is poorly known. On the other hand, method I, which is the only method that is not affected by the error in δ_{free}, still does not perform better than methods III or IV. The best performer is II, which considers the measurement of δ_{free} to be just another observation.

This effect was further investigated by making the measurement of δ_{free} of only the *host* proton imprecise. This study is summarized in Table VII. In this case,

Table VIII Relative performances of fitting methods when the standard deviation of all NMR sample observations δ_{obs} are 5 Hz, but the standard deviations of δ_{free} measurements of both protons are 0.5 Hz.

	K		D_1		D_2		
method	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
1	-2	-1	0	-4	-2	-1	-10
2	3	2	0	1	1	1	8
3	-3	-3	0	-1	-1	-3	-11
4	2	1	0	2	3	2	10
5	0	1	0	2	-1	1	3

Table IX Relative performances of fitting methods when the standard deviation of all NMR measurements are 5 Hz.

	K		D_1		D_2		
method	*med*	*sdev*	*med*	*sdev*	*med*	*sdev*	*total*
I	-2	-1	-2	-4	0	-1	-10
II	3	2	1	1	3	2	12
III	-2	-3	1	0	-2	-3	-9
IV	3	1	3	3	2	1	13
V	-2	1	-3	0	-3	1	-6

propagation of errors appears superior to method I. Method II is still superior overall, but the margin between all methods has narrowed considerably.

Two more variations in NMR observation uncertainties were studied. In these, *all* of the sample resonances δ_{obs} were assigned an uncertainty of 5 Hz.[14] In the first case, the free chemical shifts of both protons were assigned uncertainties of only 0.5 Hz; in the second case, the free chemical shifts were also assigned uncertainties of 5 Hz. The relative performances of the fitting methods under these two cases are summarized in Tables VIII and IX, respectively.

In the first of these cases, in which measurement could have given good estimates of δ_{free}, method I performs comparatively poorly. The best performances are by methods II and IV, which consider all observations but do not propagate errors. Method V, which propagates measurement errors, has somewhat intermediate performance. In the second case, in which all NMR observations are equally poor, the best performers are again methods II and IV. Methods I and III are again the worst, but the performance of method V has descended almost to their level. It should be noted that *all* methods performed poorly in these cases.

These results indicate that propagation of errors is unable to compensate for large uncertainties in δ_{free} or for large and similar measurement errors in all of the values of δ_{obs}. It is unarguably inappropriate to take a measured value of δ_{free} as the final word if that measurement is imprecise; clearly, a method such as II is then a better choice. However, NMR peak position measurements are typically *not* imprecise. Peak positions referenced to an internal standard are very reproducible. Even if a peak is broad, an assignment of its center is seldom uncertain to more than a small fraction of the peak width. Therefore, the most relevant cases to consider when evaluating fitting methods are those in which the NMR errors are negligible. In such cases, method V appears to be superior.[15]

IV. CONCLUSIONS

We have developed a method for determining the association constant K and saturation shifts D of a host/guest pair from variable-concentration NMR titration experiments. This method minimizes a sum of squared *weighted* residuals; weights are calculated by propagating measurement errors according to equations 6 and 7. Monte Carlo studies simulating realistic measurement errors and a variety of experimental designs demonstrate that this method performs well in comparison to other methods.

V. EXPERIMENTAL SECTION

A. Data Sets.

Full details of the various experimental designs are presented elsewhere.[5] We present here the range of %-bound covered by the various experiments described in Table I. Data are presented as: Expt (range % host bound; range % guest bound). H3 (1; 5-16); H4 (66-80; 2-40); H5 (28-71; 22-85); H6 (13-86; 58-99); G3 (21-35; 7-14); G4 (63-83; 17-42); G5 (14-94; 23-87); G6 (58-100; 19-86); D3 (11-31; 4-12); D4 (50-78; 20-31); D5 (64-80; 64-80); D6 (50-78; 20-31); J4 (4-64; 4-64); J5 (10-89; 10-89); J6 (46-92; 46-92); V3 (1-32; 6-9); V4 (5-81; 16-49); V5 (16-94; 23-81); V6 (73-89; 52-73).

B. Monte Carlo Comparisons.

We summarize the basic Monte Carlo procedure here – further details are provided elsewhere.[5] A Monte Carlo comparison test is carried out in the following manner:

Repeat 1500 times:

- Draw random calibration, or accuracy, errors for each delivery device used in the experiment. All volumes delivered by this device will subsequently be biased by this accuracy factor. The error for each device stays constant throughout the course of the repetition.

- Draw concentration errors for each stock solution. The simulated stock solution concentrations will differ from the intended concentrations.

- Create a sample by adding solution aliquots to the sample tube. The volume of each aliquot is determined by multiplying the appropriate delivery device's Monte Carlo accuracy factor *I* by the intended aliquot volume and then adding to this value a reproducibility error. The reproducibility error will be an absolute volume, say 0.03 μl, and is drawn afresh for each aliquot.

The sample volume is obtained by adding this aliquot volume to the volume of the sample previously in the tube. The total host and guest concentrations are determined from these volumes and from the concentrations of the combined solutions.

- Draw an error for the measurement of δ_{free} for each proton. The Monte Carlo value of δ_{free} is obtained by adding this error to the measured value.

- Draw an error for the observed chemical shift of each proton recorded in a sample. Each Monte Carlo "error-free" observation is generated by applying equation (2) to the assumed parameter values and the Monte Carlo values of $[H]_0$, $[G]_0$ and δ_{free}. To this result is added the random observation error.

- Subject the data set to analysis by each regression procedure. Each procedure generates estimates for every adjustable parameter.

After this set of 1500 replications of the experiment is complete, the medians and standard deviations of the parameter estimates from each regression procedure are calculated, and the performances of the different procedures are compared with each other. Parameter estimate medians are evaluated by their distances from the true parameter value, and parameter estimate standard deviations are evaluated by size.

Random Numbers

Uniform deviates were generated by the supplied linear congruence generator drand48 (standard C library in a Silicon graphics 4D/220 GTX workstation) and shuffled.[17] The generator was initialized at the start of each run with the value of the current system time. Normal variates were constructed from the uniform variates by the Box-Muller method,[18] and Cauchy variates were constructed from the uniform variates by a tangent transformation.[19] The generated Cauchy and normal variates were verified by the Kolmogorov-Smirnov test[20] to follow their intended theoretical distributions.

Acknowledgment:

This work was supported by the Office of Naval Research, to whom we are grateful. We also thank Professor Craig Wilcox for helpful comments.

References

1 The discussion will emphasize changes in chemical shifts of guest protons, but identical arguments can be made for host protons or other types of nuclei.
2 Connors, K.A. *Binding Constants;* Wiley: New York, **1987**.
3 Wilcox, C. S. "Design, synthesis, and evaluation of an efficacious functional group dyad. Methods and limitations in the use of NMR for measuring host-guest interactions," In *Frontiers in Supramolecular Organic Chemistry and Photochemistry,* Schneider, H.-J.; Dürr, H., Eds.; VCH: Weinheim, **1990**.
4 Barrans, R. E., Jr. manuscript in preparation.
5 For further details on many aspects of this work see: Barrans, R. E., Jr., California Institute of Technology, Ph.D. Thesis, 1993.
6 Seber, G. A. F.; Wild, C. J. *Nonlinear Regression;* Wiley: New York, 1989; Chapter 12.
7 In fact, the determination of the response variables δ_{obs} is among the most precise of all measurements performed in an NMR titration binding study.
8 For examples of studies using EMUL and MULTIFIT, see: Petti, M. A.; Shepodd, T. J.; Barrans, R. E., Jr.; Dougherty, D. A. *J. Am. Chem. Soc.*, **1988**, *110*, 6825 – 6840. Kearney, P. C.; Mizoue, L. S.; Kumpf, R. A.; Forman, J. E.; McCurdy, A.; Dougherty, D. A. *J. Am. Chem. Soc.* **1993**, *115*, 9907-9919.
9 Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences;* McGraw-Hill; New York, **1969**; p. 60.
10 The generation of these Monte Carlo data sets is described in the Experimental Section of this paper.
11 Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks;* Holden-Day: San Francisco, 1975; p. 265.
12 Creswell, C. J.; Allred, A. L. "Thermodynamic constants for hydrogen bond formation in the chloroform-benzene-cyclohexane system," *J. Phys. Chem.* **1962** *66*, 1469 – 1472.
13 Horman, I.; Dreux, B. "Estimation of Association constants of bimolecular organic complexes," *Anal. Chem.* **1983**, *55*, 1219 – 1221.
14 Assigning an uncertainty of 20 Hz to all of these measurements led to data sets that could not be adequately fit by any method.
15 The effect of non-normal measurement errors was also investigated, by considering "experimental" errors drawn from a Cauchy[16] distribution instead of from a normal distribution. As discussed in greater detail elsewhere,[5] method V still performs well unless the error distributions are especially pathological. The actual measurement errors in NMR titration studies are probably not truly normal, but they are probably closer to normality than are the Cauchy distributions considered here. If there is reason to suspect that the true measurement errors are substantially non-normal, a fitting procedure based on a more robust criterion than least-squares should be used; in such a case, none of the methods tested here would be good choices. Normal measurement errors apparently favor method V.
16 Hoel, P. G. Port, S. C.; Stone, C. J. *Introduction to Probability Theory;* Houghton Mifflin: Boston, **1971**; p. 174.
17 Press, W. H.; Flannery, D. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes: the Art of Scientific Computing;* Cambridge University: New York, **1986**; pp. 192 – 195.
18 Reference 17, p. 202.
19 Reference 16, p. 122.
20 Hoel, P. G.; Port, S. C.; Stone, C.-J. *Introduction to Statistical Theory; Houghton Mifflin: Boston,* **1971**; pp. 472 – 474.